Multi-Model Assessment of Regional Surface Temperature Trends: CMIP3 vs CMIP5 Historical (20C3M) Runs

Thomas R. Knutson, Fanrong Zeng, and Andrew T. Wittenberg

[1]Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ 08542

Version B (Draft);   July 30, 2012

To be submitted to *J. of Climate*

Email contact: Tom.Knutson@noaa.gov

To Do List:

-Citations of related work

-Later (after submission): Update analysis to compare natural forcing ensemble with all forcing ensemble, at least for the CMIP5 models. Update to have more CMIP5 models included.

Abstract.

Regional surface temperature trends as obtained from the Coupled Model Intercomparison Project 3 (CMIP3) and 5 (CMIP5) historical 20th century (20C3M) runs are compared with observed trends and with model-derived estimates of internal climate variability. We estimate the internal variability by sampling trends from the control runs of 19 CMIP3 models and 10 CMIP5 models. In the separate CMIP3 and CMIP5 analyses, we generally attempt to give different models equal weight, even when a modeling center provides fewer ensemble members or shorter control runs. The model simulations of internal climate variability are used to assess whether observed trends are “detectable” -- or highly unusual compared with internal climate variability. We also use these internal variability estimates to assess whether the simulated regional temperature trends in the various 20C3M historical runs are consistent or inconsistent with observed trends, focusing on trends of various lengths, but all ending in 2010. These tests are applied at scales ranging from global to regional -- including at any individual grid points on the observed data grid where there is sufficient data coverage over the trend period. Results are summarized using classification maps and global percent area statistics.

The CMIP3 and CMIP5 multi-model ensembles (with volcanic forcing) have warming trends that are consistent with observations over roughly 40-55% of the global area analyzed (with adequate data coverage), for trends-to-2010 that begin in start years from 1901 to 1981. The CMIP5 ensemble has about 5% higher percentage of consistent area than the CMIP3 ensemble, for trends-to-2010 that begin prior to about 1960. The fraction of analyzed global area with no detectable trend in the observations is less than 10% for trends covering 1901-2010, but this fraction gradually grows to over 50%, and is generally slightly higher for CMIP5 than CMIP3, as the trend start date shifts forward to 1991. Especially for the trends beginning earlier in the record (e.g., 1901-2010) there is some tendency for the ensemble historical run warming trend to be too large in the lower latitudes and too small at higher latitudes. The analysis identifies regions where detection of a warming trend is relatively less robust, or not detectable. These regions include (for all trend periods examined) much of the North Atlantic and North Pacific and for the more recent period (e.g., 1981-2010) the eastern tropical and subtropical Pacific and much of the extratropics in both hemispheres poleward of about 40 degrees. Conversely, the most robust warming signals, which are detectable even from trends beginning as late as 1981, generally include regions from about 40N-40S with the exception of the eastern tropical Pacific.

1. Introduction

Are historical simulations, using climate models with the best available estimates of past climate forcings, consistent with observations? This question can be examined from the viewpoint of a number of different climate variables and using different comparison methods. Here we compare modeled versus observed regional surface temperature trends, attempting to incorporate information from a large number of climate models using various multi-model combination techniques. We assess historical runs from the Coupled Model Intercomparison Project 3 (CMIP3; Meehl et al. 2007) and compare them with those from CMIP5 (Taylor et al. 2012).

The general approach used here is to compare the modeled and observed trends, in terms of both magnitude and pattern, by considering trends at each gridpoint in the observational grid, as well as trends over broader-scale regions. We use estimated internal climate variability, as simulated in the various model control runs, to assess whether observed and simulated forced trends are more extreme than those that might be expected from random sampling of internal climate variability. Similarly, we use the available ensemble of simulated forced trends to assess whether observed trends are compatible with the forcing-and-response hypotheses embodied by those forced simulations.

Formal detection/attribution techniques often use a model-generated pattern from a single or set of climate forcing experiments, and then regress this pattern against the observations to compute a scaling amplitude (e.g., Hegerl et al. 1996; Hasselmann 1997; Allen and Stott 2003) . If the scaling is significantly different from zero, the forced signal is detected. If the scaling does not significantly differ from unity, then the amplitude of the signal agrees with observations, or is at least close enough to agree within an expected range based on internal climate variability. Optimal detection techniques also filter the data during the analysis such that the chance of detecting a signal, if one is present in the data, is enhanced. In contrast to these methods, we test both the amplitude and pattern simulated in the models, and we do not apply optimization filtering to enhance prospects for detection. Our analysis is thus a consistency test for both the amplitude and pattern of the observed versus simulated trends (e.g., Knutson et al. 1999; Karoly and Wu 2005; Knutson et al. 2006). Other variants and enhancements to this general type of analysis have recently been presented by Sakaguchi et al. (2012). More discussion of various detection and attribution methods and their use in general is contained in Hegerl et al. 2009.

Our general approach in this study is to attempt to mimic observations with the models, in terms of data coverage over time. To prevent any one model from dominating the analysis, our approach attempts to weight the various models roughly equally.. Thus even if one modeling center provided ten ensemble members and another only one member, or if one center provided a much longer control run than the others, each of these models would still get an equal weighting.

(Control runs are long runs with a pre-industrial forcings that may change seasonally, but do not change from year to year.) Control runs from various modeling centers are weighted equally in the analysis, as long as the control run length is at least three times the length of the trend being examined.

In this report, the models, methods, and observed data are described in Section 2. We examine the model control runs and their variability in Section 3. Global-mean time series from the 20C3M historical runs are examined in Section 4. The grid point-based consistency tests are presented in Section 5. Section 6 contains some additional trend analysis for data averaged over larger defined regions. The discussion and conclusions are given in Section 7.

2. Model and Observed Data Sources

a. Observed data

The observed surface temperature dataset used in this study is the HadCRUT4 (Morice et al. 2012) which is available as a set of anomalies relative to the period 1961-1990. The dataset contains some notable revisions, particularly to SSTs (HadSST3; Kennedy et al. 2011) , relative to previous versions, so it important to retest earlier conclusions regarding climate trends using the revised data. The dataset also contains uncertainty information, in the form of nn-ensemble members sampling the estimated observational uncertainty.

To form a combined product of SST and land surface air temperature, Morice et al. (2012) adopt the following procedure. If both land data and SST data are available in a particular gridbox, they are weighted according to the fraction of the gridbox that is covered by land or ocean, respectively. A minimum of 25% coverage is assumed, even if the fraction of the gridbox covered by land is less than 25%. In our study, we use this same procedure to combine SST and land surface air temperature data sets from the models we analyze. .

b. CMIP3 and CMIP5 models

Figure 1 displays the complete collection of models from both CMIP3 and CMIP5 used in our analysis. The data were downloaded from the CMIP3 (www-pcmdi.gov/ipcc/about_ipcc.php) and CMIP5 (cmip-pcmdi.llnl.gov/cmip5) model archives. We regridded the model data from the 20C3M historical runs and control runs onto the observational grid. In cases where we needed to use a combined the model land surface air temperature and SST data to compare with observations, we used a procedure resembling that used for the observations, but using the model's own land-sea mask. To mimic the data gaps in the observations, we then masked out (deleted) model data at times and locations where data were labeled missing in the observations. Finally, we computed the model's climatology over the same years as for observations (1961-1990) and then created anomalies from this climatology. **[NOTE: Were 'holes' put into the Control run series? For which analyses?]**

The forcings for the 20C3M historical forcing runs are briefly summarized in **[GIVE REFERENCE SITE; SEE BAMS PAPER]**. An important distinction among the models is the treatment of volcanic forcing. 10 of the 23 CMIP3 models examined include volcanic forcing, while 13 do not. We refer to these sets of models as the "Volcanic" and "Non-Volcanic" models, respectively, and often distinguish between results for the two types of historical runs in our analysis. For cases where we include both sets, we used the term "Volc and Non-Volc" models. All 10 of the CMIP5 models included in this study included volcanic forcing.

3. Model Control Runs

a. Global mean time series

The global-mean surface air temperature series from the CMIP3 and CMIP5 model control runs are shown in Fig. 1. Data are displayed with arbitrary vertical offsets for visual clarity. The figure also shows the observed surface temperature anomalies from HadCRUT4. The curve labeled "Observed residual" was obtained by subtracting the multi-model mean of the historical volcanic forcing runs. This is an estimate of the internal variability of the climate system based on the residual from the estimated forcing response.

The control runs exhibit long-term drifts. The magnitude of these drifts tended to be larger in the CMIP3 runs than the CMIP5 control runs, although there are exceptions. We assume that these drifts are due to the models not being in equilibrium with the control run forcing, and we remove these by linear trend analysis (straight lines on figure). In some CMIP3 cases the drift proceeds at a given rate, but then the trend rate becomes smaller for the remainder of the run. We approximate the drift in these cases with two linear trend segments, as shown in the figure, which are removed to produce the drift-corrected series. The trend for these time periods is computed at each model grid point and then subtracted from the model time series. One CMIP3 model

(IAP_fgoals1.0.g) has a strong discontinuity near year 200 of the control run. We judge this as likely an artifact due to some problem with the model simulation, and we therefore chose to exclude this control run from further analysis.

None of the control runs in the CMIP3 or CMIP5 samples exhibit a centennial scale trend as large as the trend in the observations, aside from those with multi-century drifts as mentioned above. On the other hand, the variability of observed residual series appears roughly similar in scale to that from several of the control runs. Three of the CMIP3 control runs (GISS_aom, GISS_model_e_h, and GISS_model_e_f) have much lower levels of variability than in the observed residual series. The Miroc_3.2_hires model also has low variability, but the control runs is so short in length that it is used relatively little in our analysis, since we require the control run record to be at least three times as long as the trend being examined.

b. Geographical distribution of variability

The geographical distribution of the standard deviation of annual mean surface air temperature is shown in Fig. 2. for CMIP3 models and Fig. 3 for CMIP5 models. These use the full available time series from each control run. The time series have had the long-term drift removed as discussed in section (a). The features that stand out most strongly are the enhanced variability over land regions and in the eastern Equatorial Pacific. These general features (and magnitudes of standard deviation) are also seen in the observations. The observed standard deviation map is not shown here because of the relatively short observational record compared with the model control runs, and the uncertainties in removing the forced variability component from observations to create an internal variability estimate for comparison to the model control runs. Versions of the control run standard deviation map which use low pass (> X year) filtered data (not shown) indicate that most CMIP3 and CMIP5 models have their strongest low-frequency (> X year) variability in the polar regions and marginal sea ice areas near Antarctica, Greenland, and the periphery of the Arctic Ocean.

4. Global mean surface temperature: Historical runs

a. Time series of global mean surface temperature

The global mean time series of surface temperature from the 20C3M historical runs are shown in examined in Fig. 4. Thirty individual experiments using eight different models that include volcanic forcing are shown in Fig. 4 (a), while **59 experiments using 23** models (with and without volcanic forcing) are shown in (b). The model data series combines SST over oceans and surface air temperature over land, similar to observations, and masks out periods which are missing in the observed record. (All timeseries are adjusted to have zero mean in the period 1881-1920.)

The ensemble mean of the CMIP3 volcanic models (red curve in Fig. 4 (a)) agrees remarkably well with observations (black curve) although the obvious volcanically induced temporary dips are not in full agreement with the observed behavior for those periods. Nonetheless, one must consider the role of internal climate variability in judging whether these differences are significant or not. The observations are generally within the envelope of the large set of individual model simulations. The spread of the individual simulations includes the model uncertainty regarding the forced response, as well as internal variability generated by the models (e.g., Fig. 1).

The combined volcanic and non-volcanic CMIP3 runs (Fig. 4 (b)) show a substantially wider envelope of model behavior, as expected with the larger number of models and with the wider discrepancy in forcing among the models. Since the "Non-Volcanic" runs have a substantially less realistic representation of the forcing, we will generally emphasize the "Volcanic" runs in panel (a) in our forced model assessments in this study.

b. Spectra of global mean surface temperature

Figure 5 shows the spectra of observed global mean temperature and of the individual CMIP3 and CMIP5 "Volcanic forcing" historical runs from Fig. 4. The enhanced power at low frequencies is associated with the strong rising trend in both observations and models. At higher frequencies (< 10 yr periods) the model spectra are generally within the 90% confidence intervals on the observed spectral (red lines), although there is some tendency among the models for lower than observed variability levels at periods less than 10 yr.

Overall, the results of these comparisons suggest that the model simulations have a plausible representation of variability of the climate system, in terms of the spatial pattern of variability, the spectral of global mean temperature, and the direct comparison of the time series of observed and historical run global mean surface temperature. These findings encourage us to use the models to assess surface temperature trends at the regional scale in the following sections.

5. Trend assessment: detection and consistency tests

a. Global means and regional "sliding trend" analysis

In this section we compare the observed and simulated temperature trends to assess whether a particular class of systematic temperature change (linear trend) signal has emerged from the "background noise" of internal climate variability, as estimated by the models, and to assess whether the observed trends are consistent with simulated trends from the historical (20C3M) runs. We assess the trends across a wide "sliding range" of start years beginning in 1871. All trends use 2010 as the end year.

The general procedure we use is illustrated in Fig. 6 (a) for global mean temperature. The black curve in the figure shows the value of the linear trend in observed global mean temperature for each beginning year from 1871-2000 and ending in the year 2010. The trend in observed temperature is about 0.5^{o}C/100 yr early in the record but has increased to over 1.5^{o}C / 100yr by around 1980. It has decreased in recent years, being near zero since 2001. The green curve shows the "mean of ensemble means" for the **21(**?) CMIP3 climate models, where each of the **21** models is weighted equally, even if the modeling center provided a greater than average number of within-model ensemble members.

The dark blue shading in Fig. 6 (a) shows the 5th to 95th percentile range of trends for the corresponding window lengths from the long-term drift-adjusted control runs (Fig. 1). Each of 19 available CMIP3 models contributes equally to this multi-model sample, even if it has a shorter control run available. We require a control run to have at least three times the data length in question before it is included in our sampling, which is a random resampling technique across the available data. The control data was formed into 150-yr segments with random start dates for the random resampling. The 150-yr segments were then masked with the observed mask of missing data over the period 1861-2010 to create data sets with similar missing data characteristics to the observations. The analysis in Fig. 6 (a) shows that observed global temperature trends-to-2010 of almost any length are highly unusual compared to the CMIP3 simulated internal variability—even for trends as short as those beginning in 1990.

The light pink shading in Fig. 6 (a) is a measure of the uncertainty in the CMIP3 20C3M historical runs and includes the uncertainty due to different specified forcings, different forcing responses, and the influence of internal variability as simulated by the models. Under an assumption that internal variability in the control run is not substantially different from that in the forced runs, we can use the long control run for each model to estimate the component of inter-realization uncertainty that would be present in the forced trends; this is helpful, since most centers did not provide enough ensemble members to precisely assess this component of the uncertainty. The each randomly selected control run trend (used to construct the blue shading) is combined with that model's ensemble mean forced trend for that trend length, to create a distribution of historical run trends. The pink region is the 5th to 95th percentile range of this distribution of trends, and thus relates to the uncertainty of single ensemble members (which mimics the real world, itself a "single ensemble member"). In Fig. 6 (a), the black (observed) curve is always within the pink shaded region, meaning that global mean temperature trends are not obviously different from the CMIP3 historical run ensemble on any time scale, including for the most recent 'weak trends'. Therefore, for trends with starts through about 1990, the observed trend in global-mean temperature is detectable and consistent with the CMIP3 historical runs. A similar result is obtained for global mean temperature using the sample of 10 CMIP5 historical runs (Fig. 6 (b).

In contrast, when the analysis is applied to the Southeast U.S. region (Fig. 6 c, d) a much different result is obtained. The observed trend curve generally lies outside of the 5th to 95th percentile range of the forced model ensemble (pink shaded envelopes) meaning that even accounting for internal variability, the CMIP3 and CMIP5 historical runs trends-to-2010 are not consistent with the observed surface temperature trends for starting dates before about 1940. **NOTE: CHECK WITH EXPANDED VERTICAL AXIS.** Thus the CMIP3 and CMIP5 models are more easily falsified on this relatively small regional scale, meaning that there remain unexplained discrepancies between their historical simulations and observations for trends in this region.

b) Grid point-based detection and consistency assessment

The above procedure can be applied to individual gridpoints and the results displayed in map form. To do this, we create categories based on an observed trend's relation to the control run variability (e.g., pink region in Fig. 6) and its relation to the simulated historical run trends, accounting for uncertainty in the models' forced responses and internal variability. For example, if the observed trend is positive and greater than the forced response (above the pink region) we conclude that the trend is a "warming – detectable and greater than simulated". If the observed trend is positive and lies within the pink region and outside of the blue region, we conclude that the trend is "warming – detected and consistent with the simulations". If the observed trend is positive, lies below the pink region and above the blue region, we conclude that the trend is "warming- detectable but less than simulated. If the observed trend lies within the blue region, we conclude there is "no detectable change". For cooling trends, we have analogous terms to those used for the various warming cases, although these cases are relatively rare in our analysis.

In Fig. 7 (a), we show the observed surface temperature linear trend map for 1901-2010. The map shows warming at almost all locations. We assess this warming as highly unusual compared with the CMIP3 control run (internal climate) variability over most of the global region with sufficient coverage **[ADD FOOTNOTE ON SUFFICIENT COVERAGE.]**. Only in about 10% of the analyzed area (white regions in Fig. 7(c) for CMIP3 and Fig. 8(c) for CMIP5) is the trend not detectable. In a very small fraction of the analyzed area (less than 1% in either CMIP3 or CMIP5) is there a detectable cooling trend since 1901, according to our analysis.

Figure 7 (b) and 8 (b) show the multi-model ensemble trend maps for the CMIP3 and CMIP5 historical runs, weighting each of the available (volcanic) runs equally within the CMIP3 and CMIP5 analyses. We used the categorization procedure described above to categorize the observed vs. modeled trend comparison at each gridpoint (Figs. 7 (c); 8 (c). The most common categorization is of "warming-detected and consistent" (~40% of analyzed regions globally for CMIP3 and 47% for CMIP5). The second-most common categorization is of "warming – detected and greater than simulated", which is assessed for 30% (CMIP3) and 35% (CMIP5) of analyzed regions. The third-most common categorization is “warming – detected but less than simulated, which is the case for about 20% (CMIP3) and 10% (CMIP5) of the area analysed.

In Fig. 9, we show how the percent areas that we describe above change for different start years. This figure also summarizes the aggregate differences between the CMIP3 and CMIP5 results (solid lines vs. dashed lines). The percent area where the warming is detected and consistent with the CMIP3 or CMIP5 model stays consistently between about 40% and 55% for start dates ranging from 1901 to 1981. At the same time, the percent of area with no detectable change climbs steadily from 10% for 1901 start date to about 40% by 1981 start date, and reaches over 50% for 1991 start date. This illustrates the advantages of a long record for detectability of the warming trend. The increase in percent area without a detectable trend, as one slides forward in time from the 1901 start date, is compensated by a decline in the percent of area with detectable warming that is either greater than or less than simulated (i.e., outside of the ‘pink envelope’ of Fig. 6). The decline is largest for the classification “warming – detected and greater than simulated”. Comparing the CMIP3 and CMIP5 models, the two largest differences are: CMIP5 has about 5% more (~40 vs. 45%) area with detectable and consistent warming than CMIP3 for trends beginning in the first half of the 20th century, and about 10% less (~10 vs 20%) area with “warming – detected but less than simulated” for start dates from 1901 to 1931. In short, CMIP5 historical runs appear at least slightly more consistent with observed trends than the CMIP3 historical runs are, at least for the case of trends extending from the early 20th century to 2010. There is slightly less area with detectable warming trends according to the CMIP5 models, particularly for trends-to-2010 beginning from 1931 start date on.

The corresponding maps for 1951-2010 and 1981-2010 observed trends, ensemble mean historical run trends, and the categorization maps for those trends for the CMIP3 and CMIP5 models are shown in Figs. 10- 13 (panels a-c). These show the general spatial patterns associated with the changes in trend behavior for different start dates and for the CMIP3 and CMIP5 historical runs noted above. The loss of detectability, as one proceeds to mid-20th century start dates, occurs first in the extratropical North Atlantic (north of 40°N) and over large parts of the North Pacific, extending into the tropics, as seen for the 1951-2010 trends (Figs. 10 c, 11 c). For the late 20th century start dates (e.g., 1981-2010; Fig. 12c, 13c) the region of no detectable warming expands to cover most of the southern oceans, south of 40°S, and extending south from 20°S in the South Atlantic. This region also expands to include most of the eastern tropical and subtropical Pacific and much of the northern extratropics over Eurasia, North America, and the North Pacific. Tropical and subtropical regions within about 40-50 degrees of the equator (except for the eastern Pacific) are generally the regions with still a detectable (and generally consistent) warming signal, for trends beginning as late as 1981.

The remaining panels (d-n) in Figs. 7, 8, 10-13 show classification maps for the observed vs. historical runs, but in this case the metric is percentage of individual CMIP3 or CMIP5 models that are classified with the particular category for that geographic location and beginning year of the trend (all ending in 2010). That is, the determination of whether a given CMIP3 or CMIP5 individual model is included in a category (e.g., "warming- detectable and consistent") is based on the evaluation of the historical runs and control runs for that model alone. The most consistent signals across the models are for the "warming – detectable" category, which has **nine of ten or ten of ten** models in that category across large areas of the globe for 1901-2010 trends, and even for much of the tropics and subtropics for the relatively recent trends (1981-2010).

Figure 14 shows a summary statistic for the individual models. In this figure we compare the fraction of analyzed area where there is both a detectable change and where the change is consistent with the individual climate model. Note that this metric does not include the fraction of area where a climate model is consistent with observations but there is not a detectable trend. This avoids the potential "metric problem" of a model with excessive variability being consistent with a wide range of trends as an artifact of having such excessive variability.

The results in Fig. 14 show that the individual CMIP3 and CMIP5 models have rather similar behavior in terms of fraction of area with consistent detectable trends. There is somewhat more spread among the CMIP5 models (although there are more models in the sample as well.) This metric tends to reach a peak value around 1960-1970 start date before declining for later start dates.

6. Extensions and Applications of the Analysis

The analysis presented in this study introduces a framework for trend analysis that has many possible applications and extensions. Several of these, which are either planned, in progress, or completed. However, we cannot include these here as there are too many figures which do not fit within the length constraints of the journal. These extensions are briefly introduced here. We are creating a web site based largely on this analysis which will contain a growing collection of figures that will provide access to many of these extensions and applications as they become available. These are briefly discussed below.

a. Sensitivity analyses

A number of questions could be posed about our analysis, such as what do the plots look like for individual seasons, what if we had used 97th and 2.5th percentiles instead of 95th and 5th, what if we had left certain "low variability" models out of the analysis, what if we had used a different observed data set and so forth. Some of these sensitivity analyses have already been completed and are available on the above web site.

b. Focus on individual regions

Figure 15 shows a number of regions for which we have prepared extensive trend analyses like that in Fig. 6. We have done these analyses for various 4-month seasons, using CMIP3 or CMIP5 models, using 97.5th and 2.5th percentiles, leaving out certain CMIP3 control runs with lower variability levels, and other sensitivity tests. The plots are too numerous to present in this paper, but are accessible on the above web site.

c. Focus on individual models

Figures similar those in this multi-model analysis can also be prepared for individual models in the CMIP3 and CMIP5 archive. We are in the process of producing these. These analyses may be of interest as feedback to the individual centers and to others interested in individual model characteristics. The results, as they are updated, will be posted to the website above.

d. Weighting of future projections

Figure 14 shows an example of evaluation of individual models in terms of the fraction of global analyzed area with trends-to-2010 that consistent with observations. This analysis suggests a means of weighting future projections from different models based on the models' levels of agreement with past trends as in Fig. 14. We plan to explore this approach in a future study.

e. Application to Other Variables

An extension of this methodology would be explore application to other climate variables such as precipitation. We are planning to do this, beginning with precipitation, in upcoming work and to report on these developments in a future study as well as through updates and extensions of these on the above web site.

7. Summary and Conclusions

The purpose of this analysis has been to introduce and apply a framework for assessing regional surface temperature trends from the CMIP3 and CMIP5 models using a multi-model sampling approach. We showed the behavior of the various control runs of the CMIP3 and CMIP5 models. We used the control run variability to help assess whether observed trends were unusual or not compared with control run (internally generated) variability. We also used the control run variability to help assess whether observed trends were consistent with (or alternatively, significantly different from) trends from the historical (20C3M) simulations. In the separate CMIP3 and CMIP5 analyses, we generally attempt to give different models equal weight, even when a modeling center provides fewer ensemble members or shorter control runs. Test are applied at global and regional scales, as well as at individual grid points on the observed data grid where there is sufficient data coverage over the period of the trend. Results are summarized using classification maps and global percent area statistics.

Our analysis of variability (standard deviation maps, spectral analysis, and time series inspection) suggest that the CMIP3 and CMIP5 models provide a plausible representation of internal climate variability, with some likely exceptions which were noted for some models and regions.

The assessment of the trends allowed us to identify regions where the detection of warming trends is most robust (in terms of still being detectable, according to the models, for relatively late start dates, such as 1981).   These regions tend to be in the tropics and subtropics, but outside of the eastern Pacific, which is influenced by strong interannual variability associated with ENSO. The reduced global area with detectable trends as one examines later start dates for trends in the record (all trends ending in 2010) illustrates the advantages of long records for trend detection in the context of this model-based assessment. The analysis also suggests a modestly closer agreement of models with observed trends for CMIP5 models compared to CMIP3 models—at least for the relatively longer trends-to-2010 that begin in the first half of the 20th century.

For trends-to-2010 beginning from the early 20th century, about 40-50% of the analyzed regions globally have a detectable warming that is consistent with the 20C3M historical runs, with slightly higher percentage for the CMIP5 simulations. The fraction of area with no detectable change is only about 10% for trends 1901-2010, but increases steadily to over 50% as the beginning year is moved forward to 1981. The fraction of area with detectable and consistent warming stays relatively constant for start years through about 1981, before falling below 40% for trends from 1991-2010. The "loss" of detectable warming regions as one moves forward with the start dates, is mainly a "loss" in regions with detectable warming that is inconsistent with the historical runs, which decreases from about 50% for 1901-2010 to less than 10% for trends 1991-2010. That is, for the most recent trends (1991-2010), the trends are classified predominantly as either non-detectable relative to the control runs, or as detectable warming that is consistent with model historical runs (for both CMIP3 and CMIP5 models). The shorter the epoch, the larger the contribution of internal variability to the trend, leading to a greater spread (uncertainty) for sampled trends.

As has been noted in a previous paper using a similar methodology with two climate models (Knutson et al. 2006), disagreement between modeled and observed trends in this type of analysis can occur due to shortcomings of models (internal variability simulation; response to forcing), shortcomings of the specified specified historical forcings, or problems with the observed data. The HadCRUT4 data set **{REF]E]**contains multiple ensemble members that attempt to characterize the uncertainties in the observations. We have performed tests on these ensembles to assess the spread of observed trend estimates. These tests thus far indicate that even at the regional scale, the spread in trend estimates due to observational uncertainties, as contained in the ensembles, is generally much smaller than the spread in model simulated trends due to both internal variability and differences in forced responses in the historical runs (e.g., Fig. 6). However, it is possible that other observational datasets could have somewhat different trends.

We have attempted to at least partially address the issue of uncertainties in the simulation of internal climate variability and in the response to historical forcing by using multi-model ensembles. Nonetheless, the CMIP3 and CMIP5 simulations represent an “ensemble of opportunity” which cannot necessarily be expected to represent the true structural uncertainty in results, due to shortcomings/uncertainties in the models and climate forcings.

While these issues lack a final resolution, the methodology shown here can at least help to quantify the uncertainties associated with the climate change detection problem. The results show that when CMIP3 and CMIP5 historical runs are confronted with observed surface temperature variations and trends, across a wide range of trend start dates and at various geographical locations around the globe, warming is found that is generally consistent with forced simulations but not with unforced simulations. This provides further support for the claim of a discernable influence of humans on climate, via anthropogenic forcing agents like increased greenhouse gases. A future enhancement of these findings would be to compare the CMIP5 all-forcing historical runs with runs that include only natural forcings, to provide a more direct assessment of the roles of anthropogenic versus natural forcings in observed temperature trends at the regional scale.

Acknowledgements. To be completed.

References

Allen, M. R., and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting. Part I: Theory. Clim. Dyn., 21, 477-491.

Allen, M. R., and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting. Clim. Dyn., 15, 419-434.

Hasselmann, K., 1997: Multi-pattern fingerprint method for detection and attribution of climate change. Clim. Dyn., 13, 601-612.

Hegerl, G. C., et al. 2009: Good practice guidance paper on detection and attribution related to anthropogenic climate change. Available from IPCC: www.ipcc.ch/pdf/supporting-material/ipcc_good_practice_guidance_paper_anthropogenic.pdf

Hegerl, G.C., H. v. Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996: Detecting greenhouse gas induced climate change with an optimal fingerprint method. J. Climate, 9, 2281-2306.

Karoly, D.J., and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Clim.*, **18**, 4337–4343.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, part 2: biases and homogenization. J. Geophys. Res., 116, D14104, doi:10.1029/2010JD015220.

Knutson, T.R., T.L. Delworth, K.W. Dixon, and R.J. Stouffer, 1999: Model assessment of regional surface temperature trends (1949-1997). *J.Geophys. Res.*, **104**, 30981–30996.

Knutson, T.R., et al., 2006: Assessment of twentieth-century regional surface temperature trends using the GFDL CM2 coupled models. *J. Clim.*, **19**, 1624–1651.

Meehl, G. A. et al., 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.* **88,** 1383–1394.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observation estimates: The HadCRUT5 data set. *J. Geophys. Res.*, 117, D08101, doi:10.1029/2011JD017187.

Sakaguchi, K. X. Zeng, and M. A. Brunke, 2012: Temporal- and Spatial-scale dependence of three CMIP3 climate models in simulating the surface temperature trend in the twentieth century. J. Climate, 25, 2456-2470.

Santer, B. D., T. M. L. Wigley, and P. D. Jones, 1993: Correlation method in fingerprint detection studies. Clim. Dyn., 8, 265-276.

Schneider, T., and I.M. Held, 2001: Discriminants of twentieth-century changes in Earth surface temperatures. *J. Clim.*, **14**, 249–254.

Shin, S.-I., and P. D. Sardeshmuhk, 2011: Critical influence of the pattern of tropical ocean warming on remote climate trends. Clim. Dyn., 36, 1577-1591.

Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93,** 485-498.

Figure Captions

Fig. 1. Timeseries of global mean annual mean surface air temperature anomalies from the CMIP3 (a, b) and CMIP5 (c) preindustrial control runs. Observed global mean surface temperature (HadCRUT4, combining SST and land surface air temperature anomalies) is also shown on the diagrams for comparison. The curves labeled "Observed residual" or "HadCRU4 residual" were created by subtracting the multi-model ensemble mean surface temperature (from masked SSTs and land surface air temperatures from the 20C3M historical runs for either CMIP3 or CMIP5) from the observed temperature. Straight lines (one or two segments) through the control run time series depict the long term linear drift. The long term drift over these years is calculated at each grid point and then subtracted from the model control run series before performing further analysis in our study. The various curves have been displaced vertically by arbitrary constants for visual clarity.

Fig. 2. Standard deviation (^{o}C) of annual mean surface air temperature from the CMIP3 pre-industrial control runs (e.g., Fig. 1 a,b). The long term linear drifts (periods identified by the linear line segments in Fig. 1 a,b) were removed prior to computing the standard deviation.

Fig. 3. As in Fig. 2 but for the 10 CMIP5 models analyzed in this study.

Fig. 4. Timeseries of global mean surface temperature anomalies (combined SST and land surface air temperature) from observations (HadCRUT4; black curves) and CMIP3 (a, b) or CMIP5 (c) 20C3M historical runs (orange curves) in degrees Celsius. The historical runs in (b) include CMIP3 models with and without volcanic forcing. Those in (a) are from CMIP3 models with volcanic forcing. All of the CMIP5 model runs shown in (c) included volcanic forcing. The red curves show the multi-model ensemble means, which was computed by weighting each model equally (as opposed to each individual model run equally). All series have been re-centered so that the mean value for the years 1881-1920 is zero. Model data were masked with the observed temporally evolving missing data mask.

Fig. 5. Variance spectra as a function of frequency for observed global mean surface temperature (combined SST and land surface air temperature) plotted against that for the individual (a) CMIP3 and (b) CMIP5 "Volcanic forcing" historical runs from Fig. 4. The spectra

in (c) and (d) are based on observed or model historical runs where the multi-model ensemble surface temperature from the 20C3M historical runs is subtracted from the observed global mean temperature series to form a residual. Similarly, the multi-model ensemble is subtracted from each individual historical run to form a modeled residual for comparison to the observed.

Fig. 6. Trends (deg C/100 yr) in surface temperature as a function of starting year, with all trends ending in 2010, for the CMIP3 (a,c) and CMIP5 (b,d) models. The black curves are from observations (HadCRUT4). The green curves are the multi-model ensembles, with each model weighted equally. The blue shading shows the 5th to 95th percentile range of trends of the given length based on random resampling of the model control runs, with each model sampled equally frequently regardless of control run length. The pink shading shows the range obtained by using the same control run samples as for the blue shading, but adding onto each control run trend the ensemble mean trend, from the given start year, of that model's all forcing run. Violet shading shows where the pink and blue shaded regions overlap. Region used: Global (a,c) or the Southeast United States (b,d), with boundaries of the latter region shown in Fig. 15.

Fig. 7. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP3 multi-model (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of the multi-model ensemble mean trends and variability in (c). In (c) the darkest red color shows regions where a detectable warming occurred, but is significantly greater than simulated (see text); lighter red shows regions with detectable warming that is consistent with observations; orange shows regions with a warming that is detectable, but less than simulated; white areas show regions with no detectable trend; light blue shows regions with cooling that is detectable but less than simulated; medium blue shows regions with detectable cooling that is consistent with the models; and dark blue shows regions with detectable cooling that is cooler than simulated in the historical runs. Panels (d-h) show the fraction of the 10 individual CMIP3 models whose historical forcing (including volcanic) runs meet the criteria listed below the panel. The criteria are: d) detectable cooling that is more than simulated; e) detectable cooling that is consistent with the model; f) detectable cooling that is less than simulated; g) no detectable change; h) detectable warming that is less than simulated; i) detectable warming that is consistent with the model; j) detectable warming that is more than simulated; k) detectable warming (sum of h,i,j); l) detectable warming that is consistent or greater than simulated (i+j); m) observed and simulated trends are consistent (including non-detectable changes that are consistent); and n) observed and simulated trends are inconsistent (1-m).

Fig. 8. As in Fig. 7, but for the ten CMIP5 models analyzed in the study.

Fig. 9. Summary assessment of trends-to-2010 comparing the CMIP3 (solid lines) and CMIP5 (dashed lines) multi-model ensembles (historical 20C3M runs with volcanic forcing). The fraction of global analyzed areas meeting certain criteria (see graph labels) are shown as a function of start year.

Fig. 10. As in Fig. 7, but for trends over the period 1951-2010.

Fig. 11. As in Fig. 7, but trends over the period 1951-2010 for the ten CMIP5 models analyzed in the study.

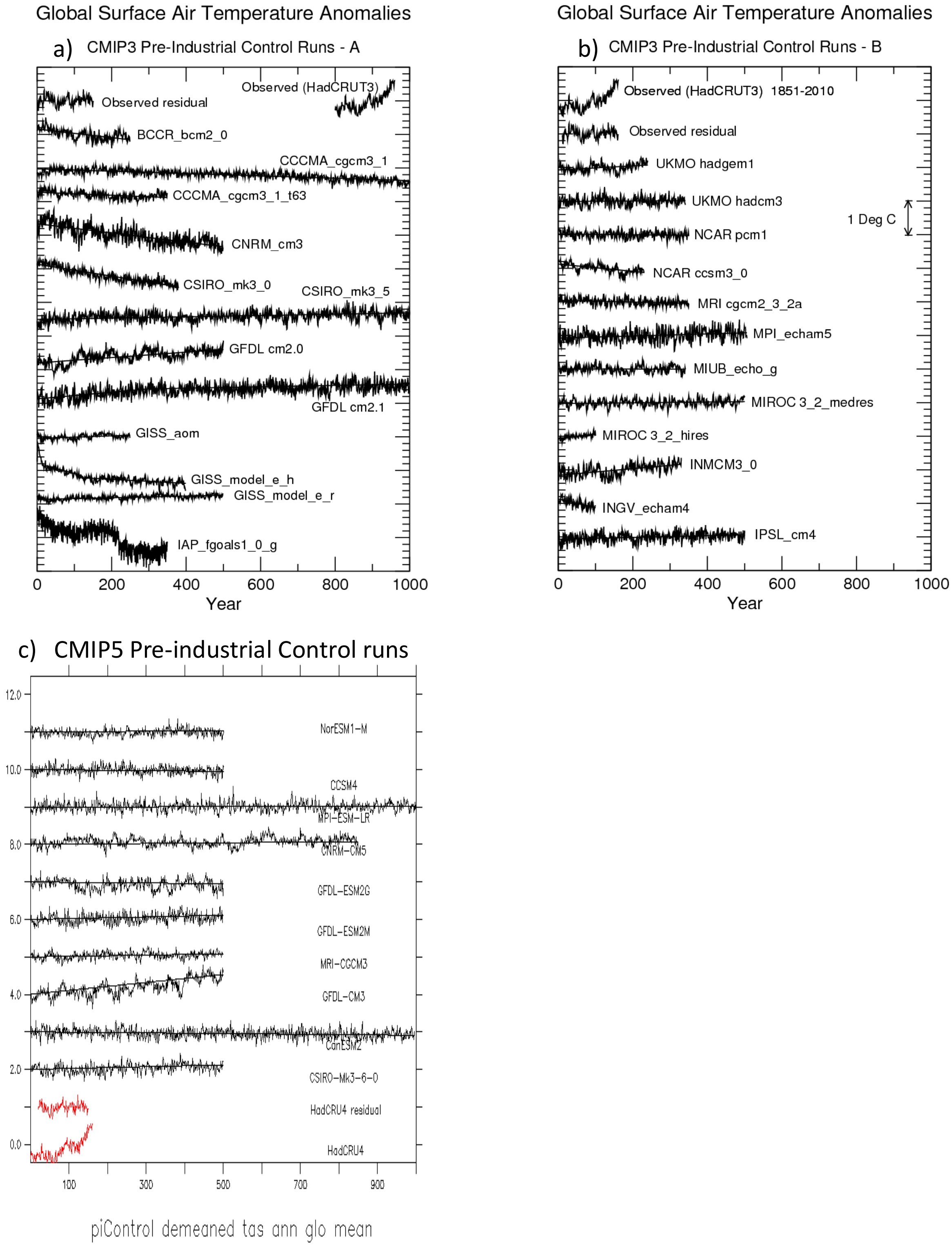
Fig. 12. As in Fig. 7, but for trends over the period 1981-2010.

Fig. 13. As in Fig. 7, but trends over the period 1981-2010 for the ten CMIP5 models analyzed in the study.

Fig. 14. Individual a) CMIP3 and b) CMIP5 models are assessed for consistency with observed surface temperature trends-to-2010 for start years from 1901 to 1991. Plotted is the percent of analyzed global area where each model's (legend) multi-member ensemble mean trends are consistent (accounting for internal variability) with the observed trends. The trends are analyzed at each grid point where there is sufficient temporal data coverage for the trend in question (see text).

Fig. 15. Map illustrating regions where trend analyses (like those in Fig. 6, but with additional augmented analyses as discussed in the text) are available online (web site).

Fig. 1



Global Surface Air Temperature Anomalies

a) CMIP3 Pre-Industrial Control Runs - A

Global Surface Air Temperature Anomalies

b) CMIP3 Pre-Industrial Control Runs - B

c) CMIP5 Pre-industrial Control runs

Fig. 2

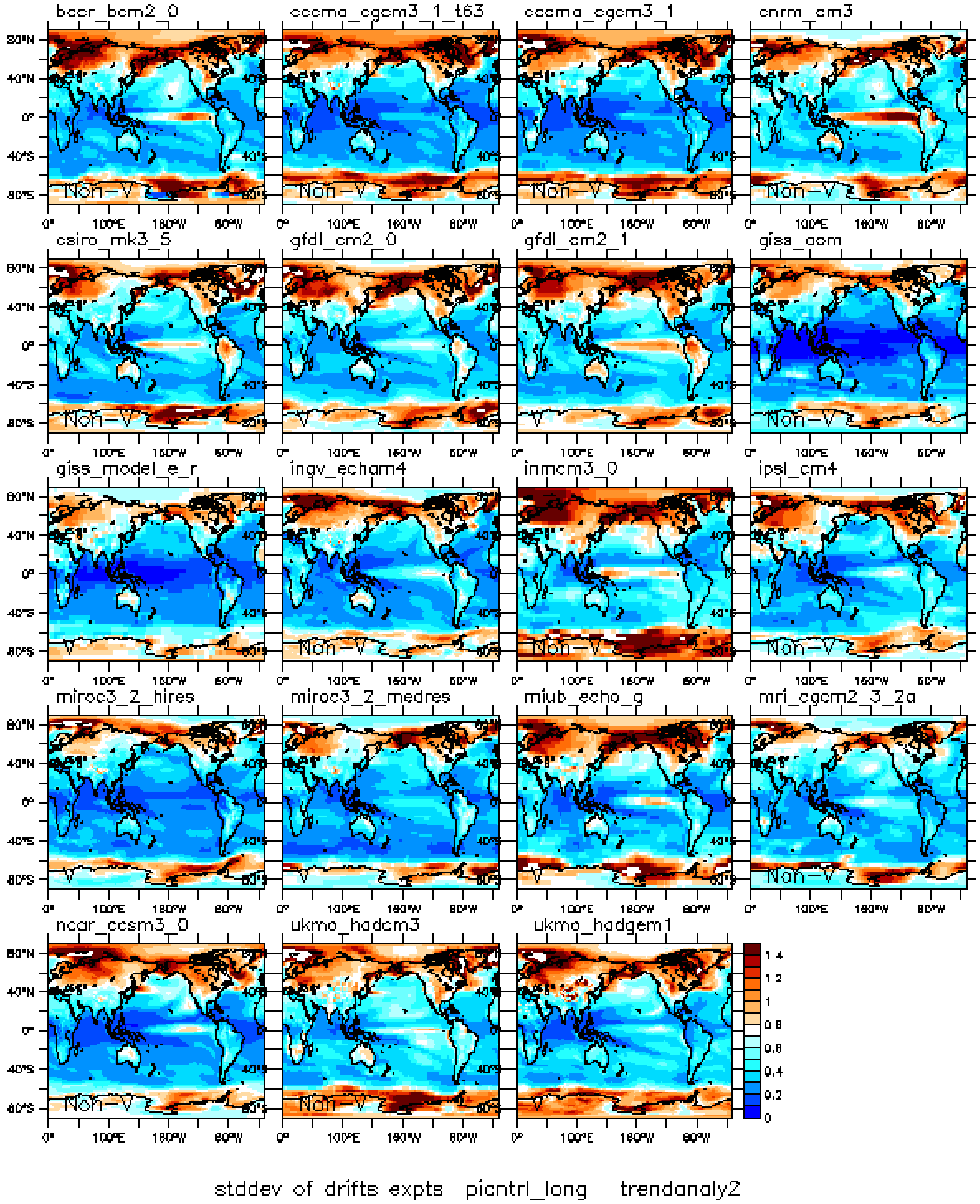
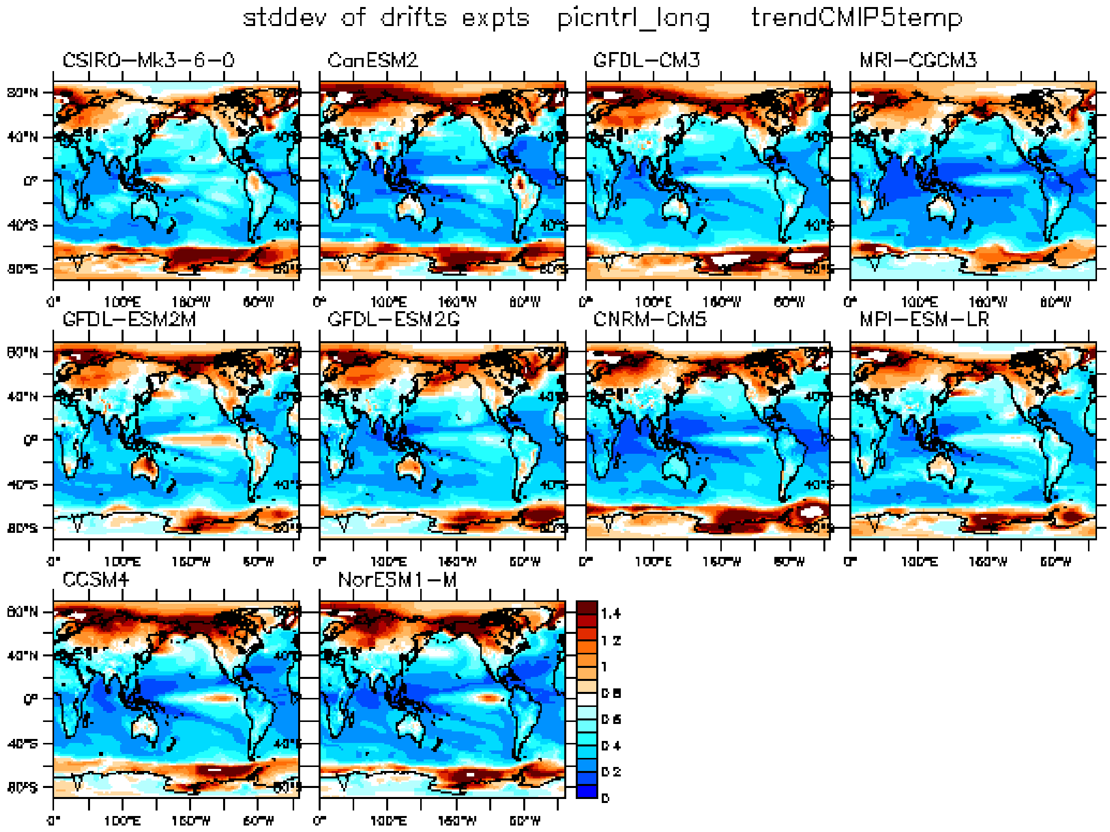

bccr_bcm2_0
cccma_cgcm3_1_t63
cccma_cgcm3_1
cnrm_cm3
csiro_mk3_5
gfdl_cm2_0
gfdl_cm2_1
giss_aom
giss_model_e_r
ingv_echam4
inmcm3_0
ipsl_cm4
miroc3_2_hires
miroc3_2_medres
miub_echo_g
mri_cgcm2_3_2a
ncar_ccsm3_0
ukmo_hadcm3
ukmo_hadgem1
Non-V
V
80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
1.4
1.2
1
0.8
0.6
0.4
0.2
0
stddev of drifts expts picntrl_long trendanaly2

Fig. 3



stddev of drifts expts picntrl_long trendCMIP5temp

CSIRO−Mk3−6−0

CanESM2

GFDL−CM3

MRI−CGCM3

GFDL−ESM2M

GFDL−ESM2G

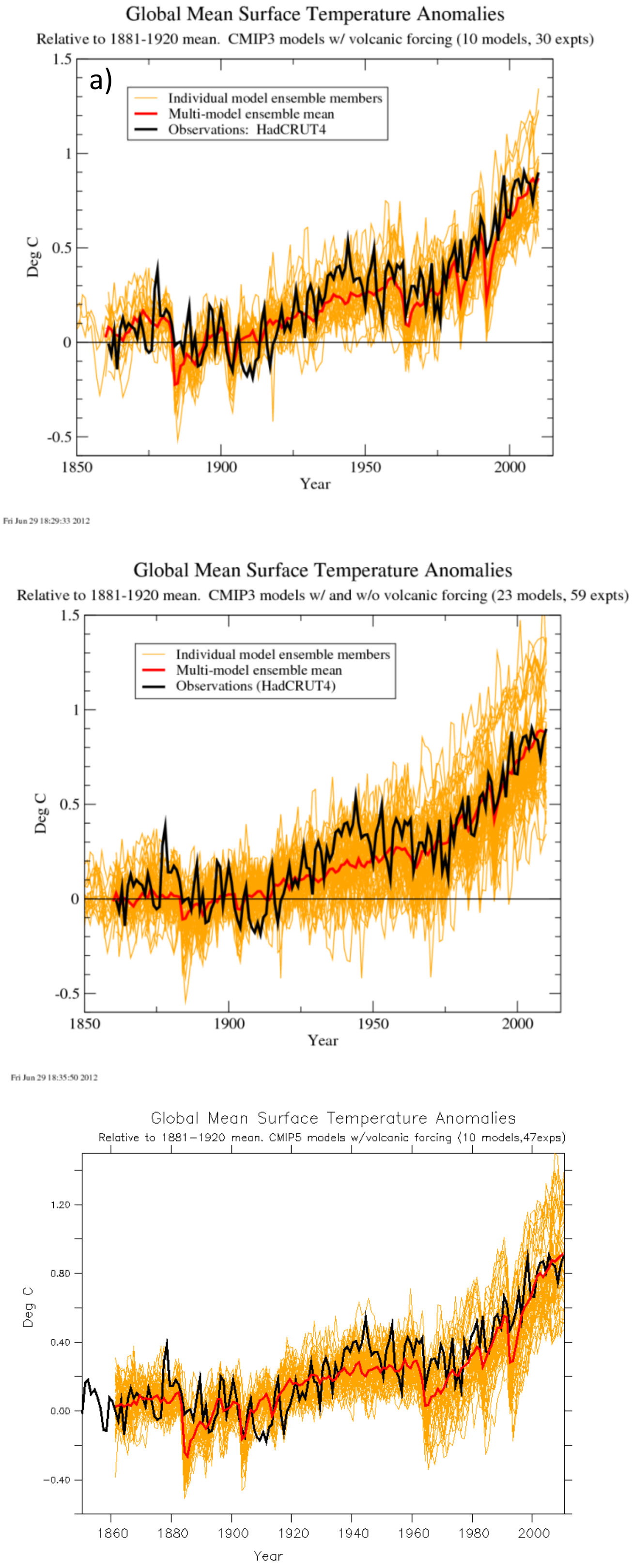
CNRM−CM5

MPI−ESM−LR

CCSM4

NorESM1−M

Fig. 4



Global Mean Surface Temperature Anomalies

Relative to 1881-1920 mean. CMIP3 models w/ volcanic forcing (10 models, 30 expts)

Fri Jun 29 18:29:33 2012

Global Mean Surface Temperature Anomalies

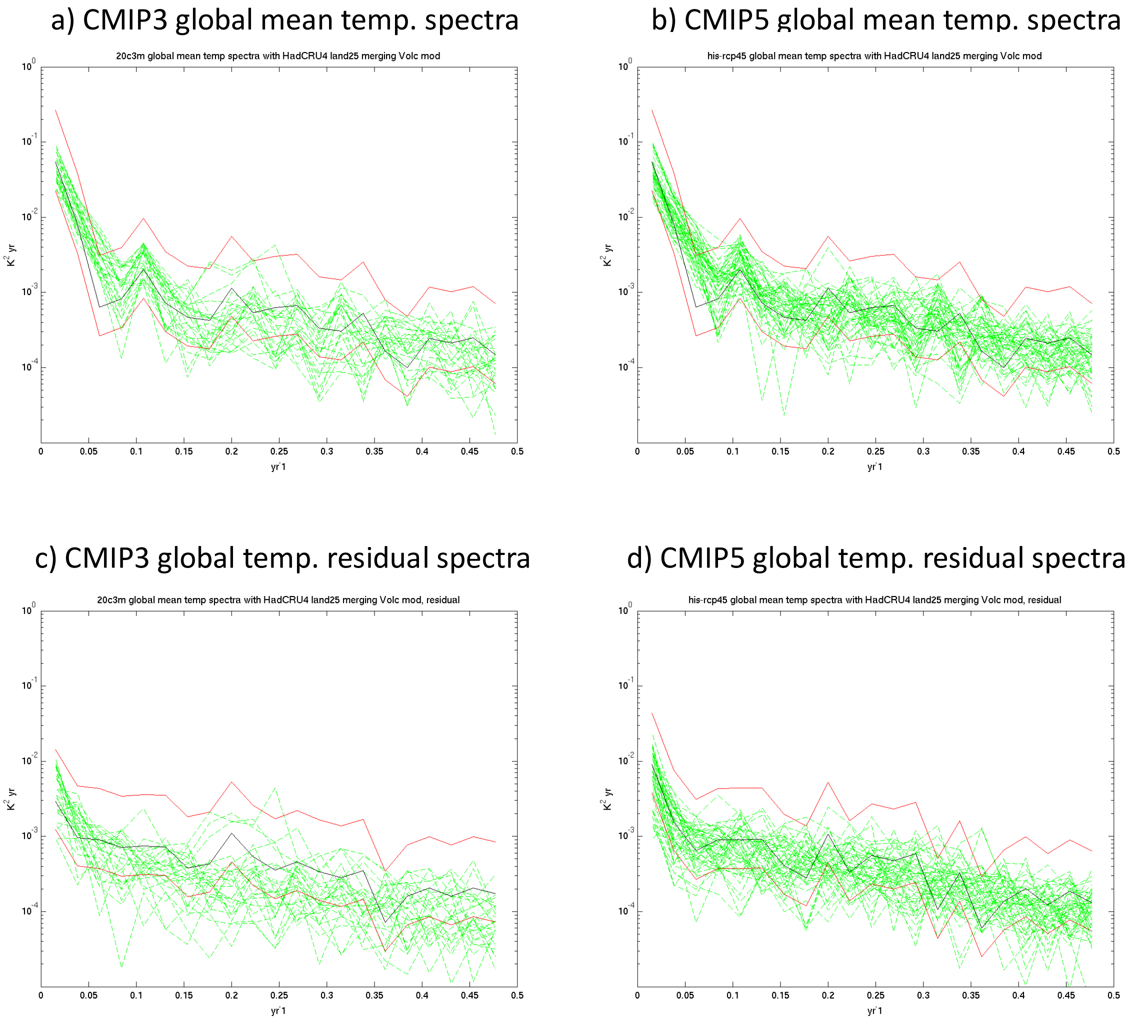
Relative to 1881-1920 mean. CMIP3 models w/ and w/o volcanic forcing (23 models, 59 expts)

Fri Jun 29 18:35:50 2012

Global Mean Surface Temperature Anomalies

Relative to 1881−1920 mean. CMIP5 models w/volcanic forcing (10 models,47exps)

Fig. 5



a) CMIP3 global mean temp. spectra

b) CMIP5 global mean temp. spectra

c) CMIP3 global temp. residual spectra

d) CMIP5 global temp. residual spectra

Fig. 6

Note: adjust vertical axis scaling

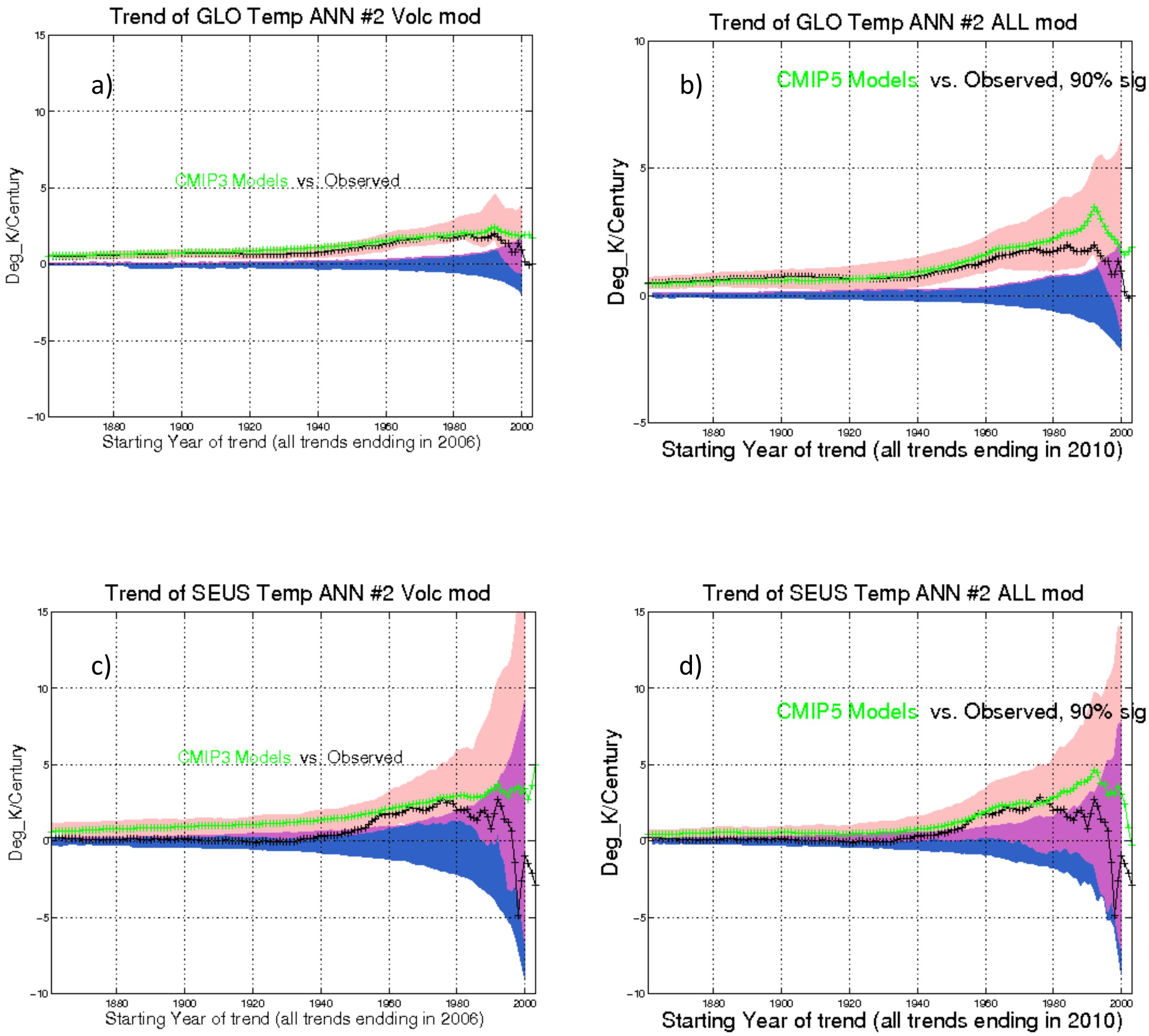


Fig. 7

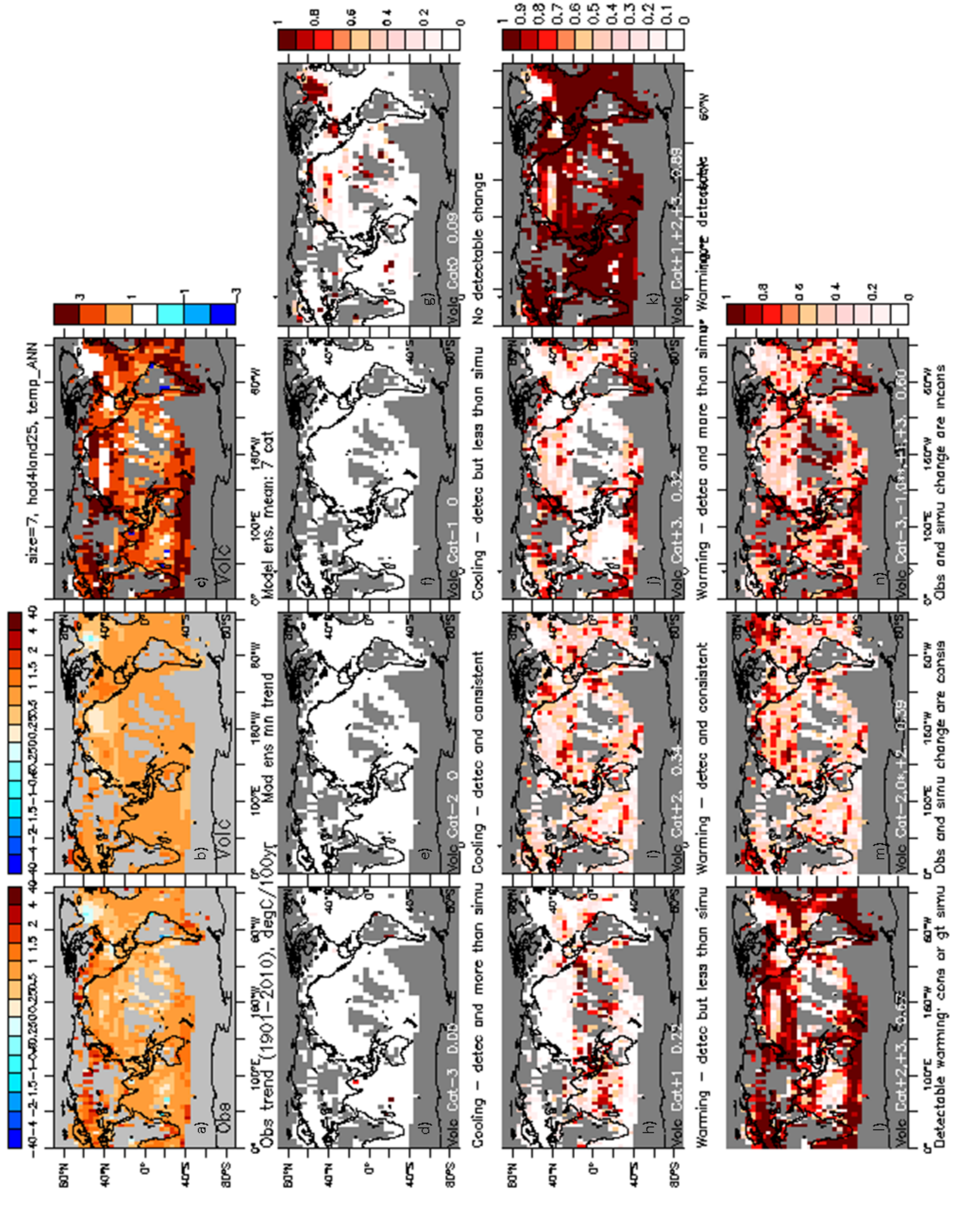

size=7, had4land25, temp_ANN
Obs trend (1901–2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
a) Obs
b) Volc
c) Volc
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
-3 -1 1 3
80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
d) Volc Cat-3 0.05
Cooling - detec and more than simu
e) Volc Cat-2 0
Cooling - detec and consistent
f) Volc Cat-1 0
Cooling - detec but less than simu
g) Volc Cat0 0.09
No detectable change
0 0.2 0.4 0.6 0.8 1
h) Volc Cat+1 0.24
Warming - detec but less than simu
i) Volc Cat+2 0.34
Warming - detec and consistent
j) Volc Cat+3 0.32
Warming - detec and more than simu
k) Volc Cat+1,+2,+3 0.89
Warming - detectable
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
l) Volc Cat+2,+3 0.67
Detectable warming' cons or gt simu
m) Volc Cat-2,0*,+2 0.39
Obs and simu change are consis
n) Volc Cat-3,-1,0*,+1,+3 0.60
Obs and simu change are incons
0 0.2 0.4 0.6 0.8 1

Fig. 8

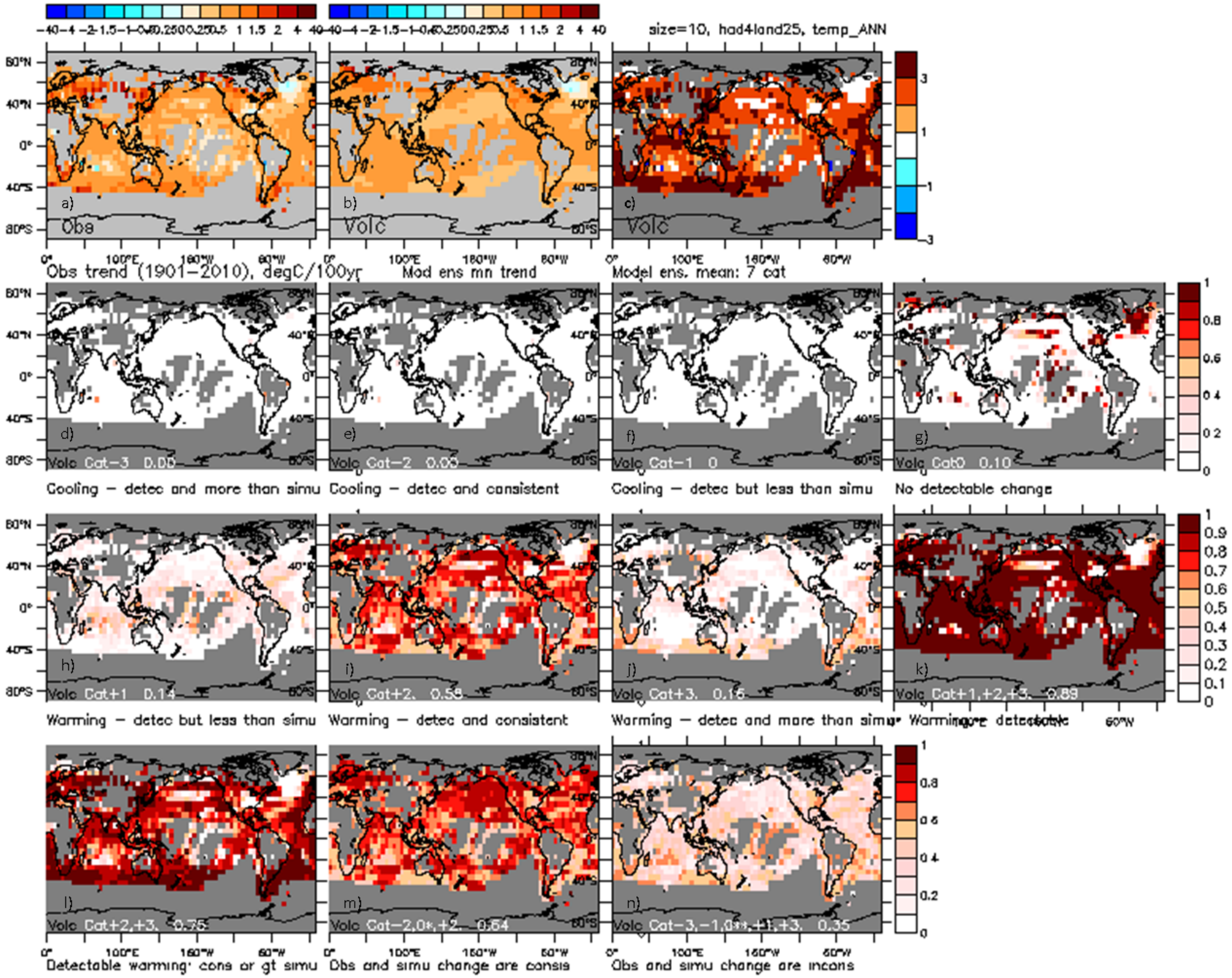

size=10, had4land25, temp_ANN
a)
Obs
b)
Volc
c)
Volc
Obs trend (1901–2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
d)
Volc Cat-3 0.00
e)
Volc Cat-2 0.00
f)
Volc Cat-1 0
g)
Volc Cat0 0.10
Cooling – detec and more than simu
Cooling – detec and consistent
Cooling – detec but less than simu
No detectable change
h)
Volc Cat+1 0.14
i)
Volc Cat+2. 0.58
j)
Volc Cat+3. 0.16
k)
Volc Cat+1,+2,+3. 0.89
Warming – detec but less than simu
Warming – detec and consistent
Warming – detec and more than simu
Warming detectable
l)
Volc Cat+2,+3. 0.75
m)
Volc Cat-2,0*,+2. 0.64
n)
Volc Cat-3,-1,0**,+1,+3. 0.35
Detectable warming' cons or gt simu
Obs and simu change are consis
Obs and simu change are incons
80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
3
1
-1
-3
1
0.8
0.6
0.4
0.2
0
0.9
0.7
0.5
0.3
0.1

Fig. 9.

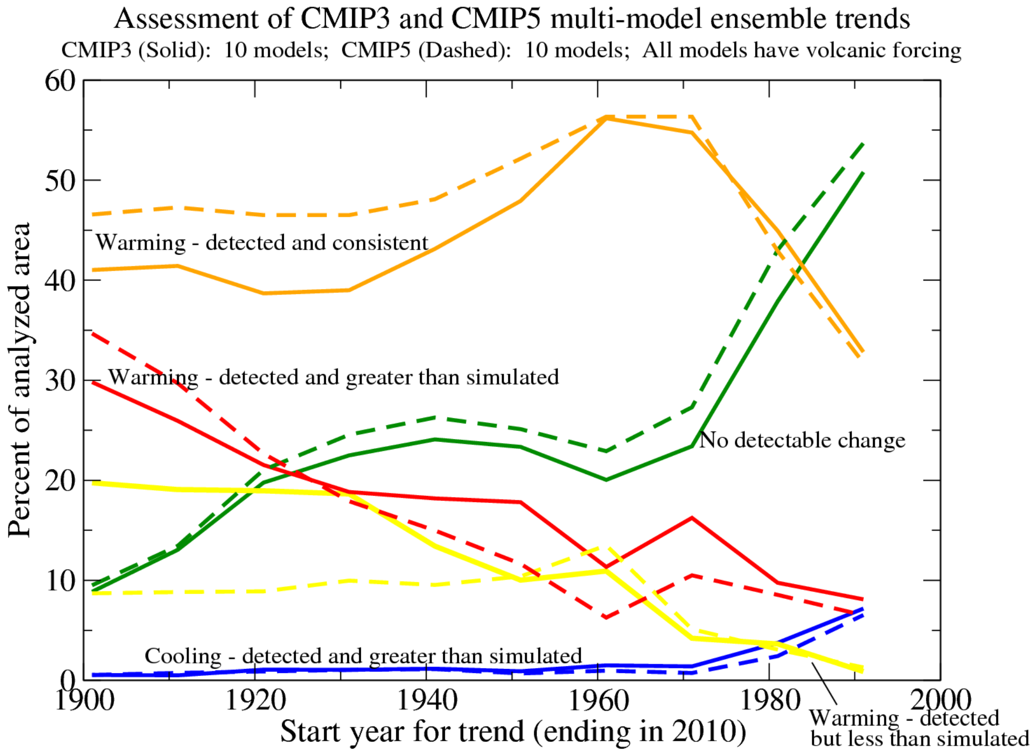

Assessment of CMIP3 and CMIP5 multi-model ensemble trends
CMIP3 (Solid): 10 models; CMIP5 (Dashed): 10 models; All models have volcanic forcing
Percent of analyzed area
0
10
20
30
40
50
60
1900
1920
1940
1960
1980
2000
Start year for trend (ending in 2010)
Warming - detected and consistent
Warming - detected and greater than simulated
No detectable change
Cooling - detected and greater than simulated
Warming - detected but less than simulated

Fig. 10

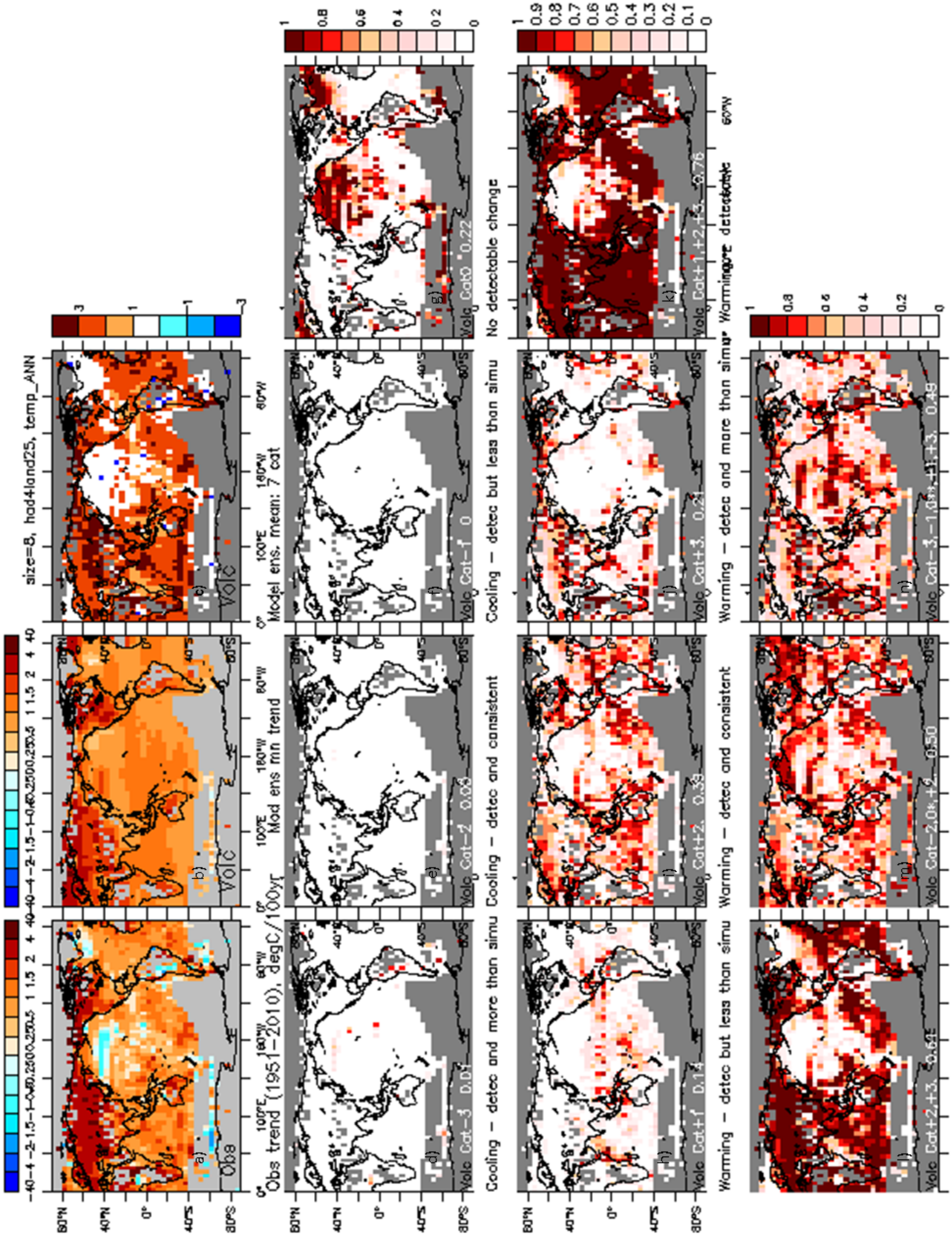

size=8, had4land25, temp_ANN
Obs trend (1951-2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
3
1
-1
-3
1
0.8
0.6
0.4
0.2
0
0.9
0.7
0.5
0.3
0.1
80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
a)
Obs
b)
Volc
c)
Volc
d)
Volc Cat-3 0.01
e)
Volc Cat-2 0.03
f)
Volc Cat-1 0
g)
Volc Cat0 0.22
h)
Volc Cat+1 0.14
i)
Volc Cat+2 0.39
j)
Volc Cat+3 0.21
k)
Volc Cat+1,+2,+3 0.76
l)
Volc Cat+2,+3 0.61
m)
Volc Cat-2,0*,+2 0.50
n)
Volc Cat-3,-1,0*,+1,+3 0.49
Cooling - detec and more than simu
Cooling - detec and consistent
Cooling - detec but less than simu
No detectable change
Warming - detec but less than simu
Warming - detec and consistent
Warming - detec and more than simu
Warming' more detectable
Detectable warming' cons or gt simu
Obs and simu change are consis
Obs and simu change are incons

Fig. 11

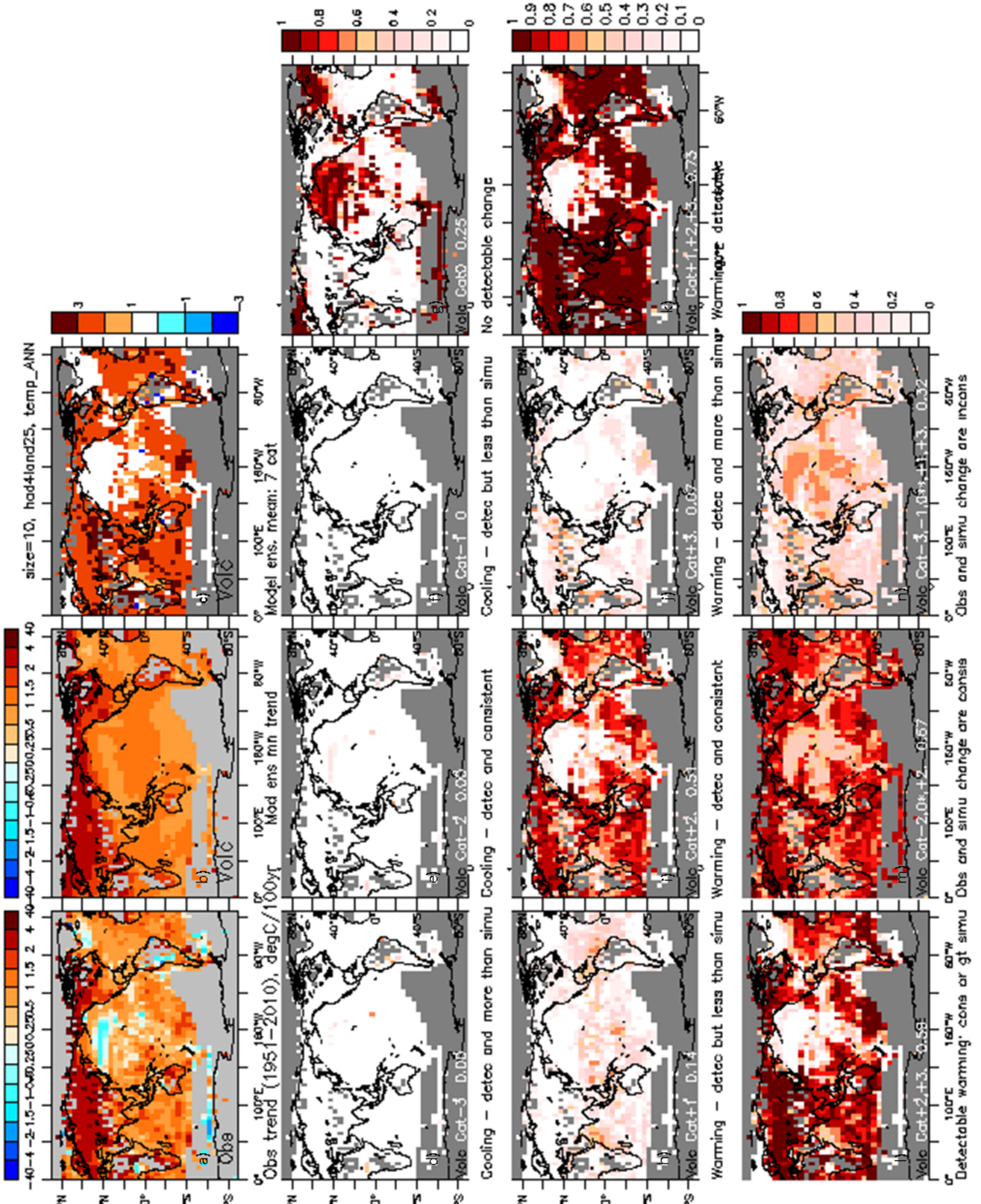

size=10, had4land25, temp_ANN
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
3 1 -1 -3
a) Obs
b) Volc
c) Volc
Obs trend (1951-2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
d) Volc Cat-3 0.00
Cooling - detec and more than simu
e) Volc Cat-2 0.00
Cooling - detec and consistent
f) Volc Cat-1 0
Cooling - detec but less than simu
g) Volc Cat0 0.25
No detectable change
1 0.8 0.6 0.4 0.2 0
h) Volc Cat+1 0.14
Warming - detec but less than simu
i) Volc Cat+2 0.51
Warming - detec and consistent
j) Volc Cat+3 0.07
Warming - detec and more than simu
k) Volc Cat+1,+2,+3 0.73
Warming more detectable
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
l) Volc Cat+2,+3 0.58
Detectable warming' cons or gt simu
m) Volc Cat-2,0,+2 0.67
Obs and simu change are consis
n) Volc Cat-3,-1,0,+1,+3 0.32
Obs and simu change are incons
1 0.8 0.6 0.4 0.2 0
80°N 40°N 0° 40°S 80°S
0° 100°E 160°W 60°W

Fig. 12

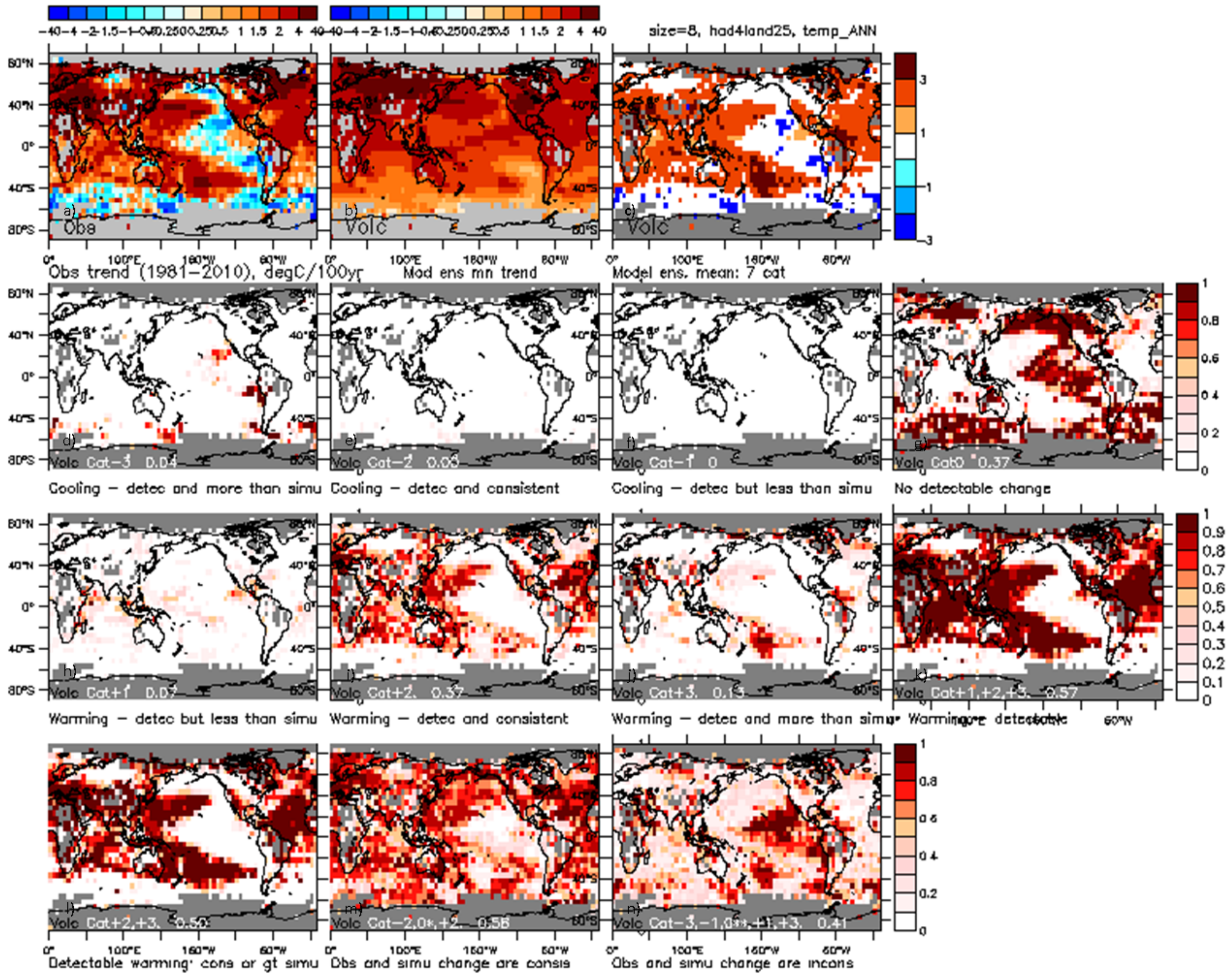

size=8, had4land25, temp_ANN
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
a) Obs
b) Volc
c) Volc
Obs trend (1981–2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
3
1
-1
-3
80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
d) Volc Cat-3 0.04
e) Volc Cat-2 0.03
f) Volc Cat-1 0
g) Volc Cat0 0.37
Cooling – detec and more than simu
Cooling – detec and consistent
Cooling – detec but less than simu
No detectable change
h) Volc Cat+1 0.07
i) Volc Cat+2. 0.37
j) Volc Cat+3. 0.13
k) Volc Cat+1,+2,+3. 0.57
Warming – detec but less than simu
Warming – detec and consistent
Warming – detec and more than simu
Warming – detectable
l) Volc Cat+2,+3. 0.50
m) Volc Cat-2,0*,+2. 0.56
n) Volc Cat-3,-1,0**,+1,+3. 0.41
Detectable warming: cons or gt simu
Obs and simu change are consis
Obs and simu change are incons
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig. 13

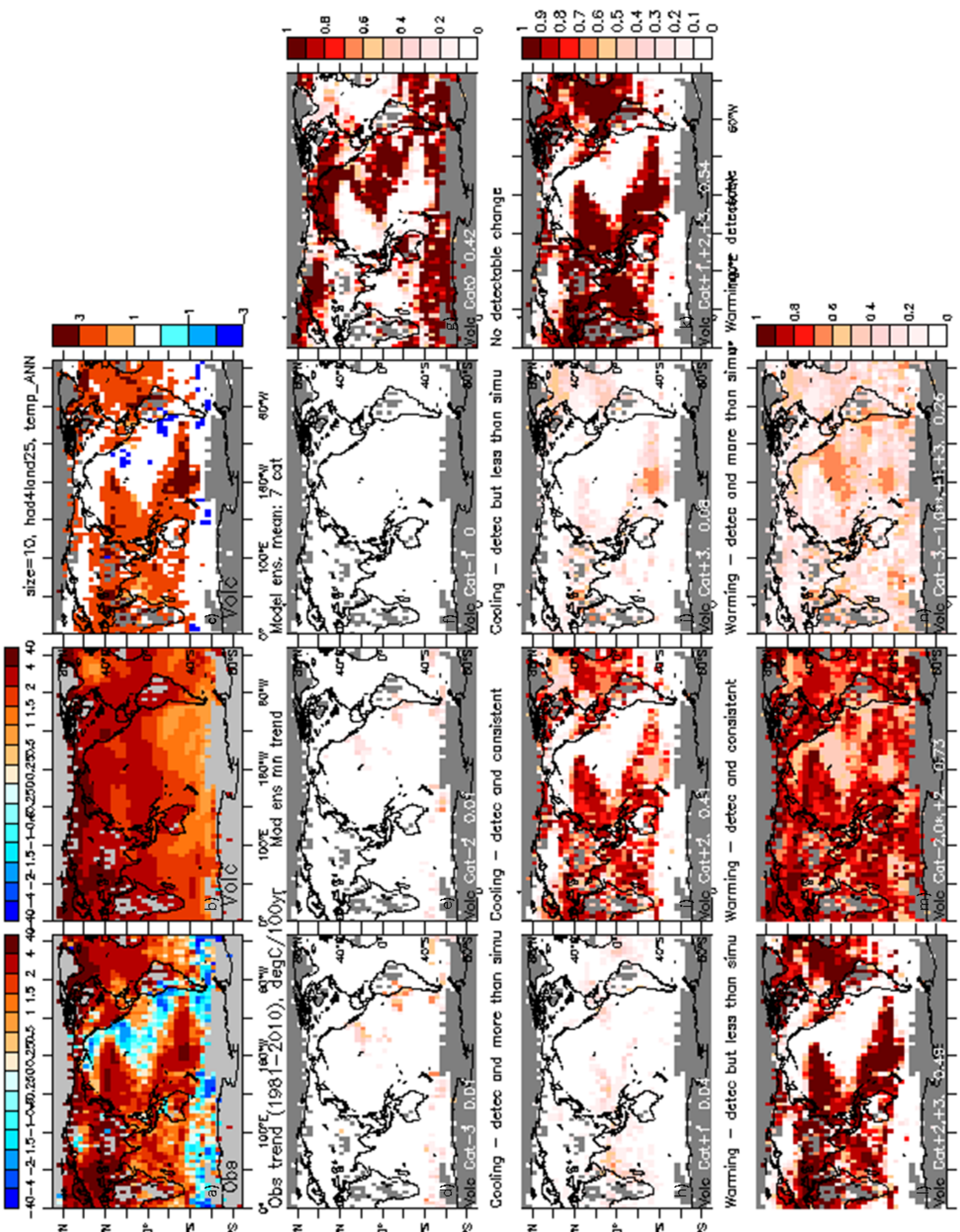
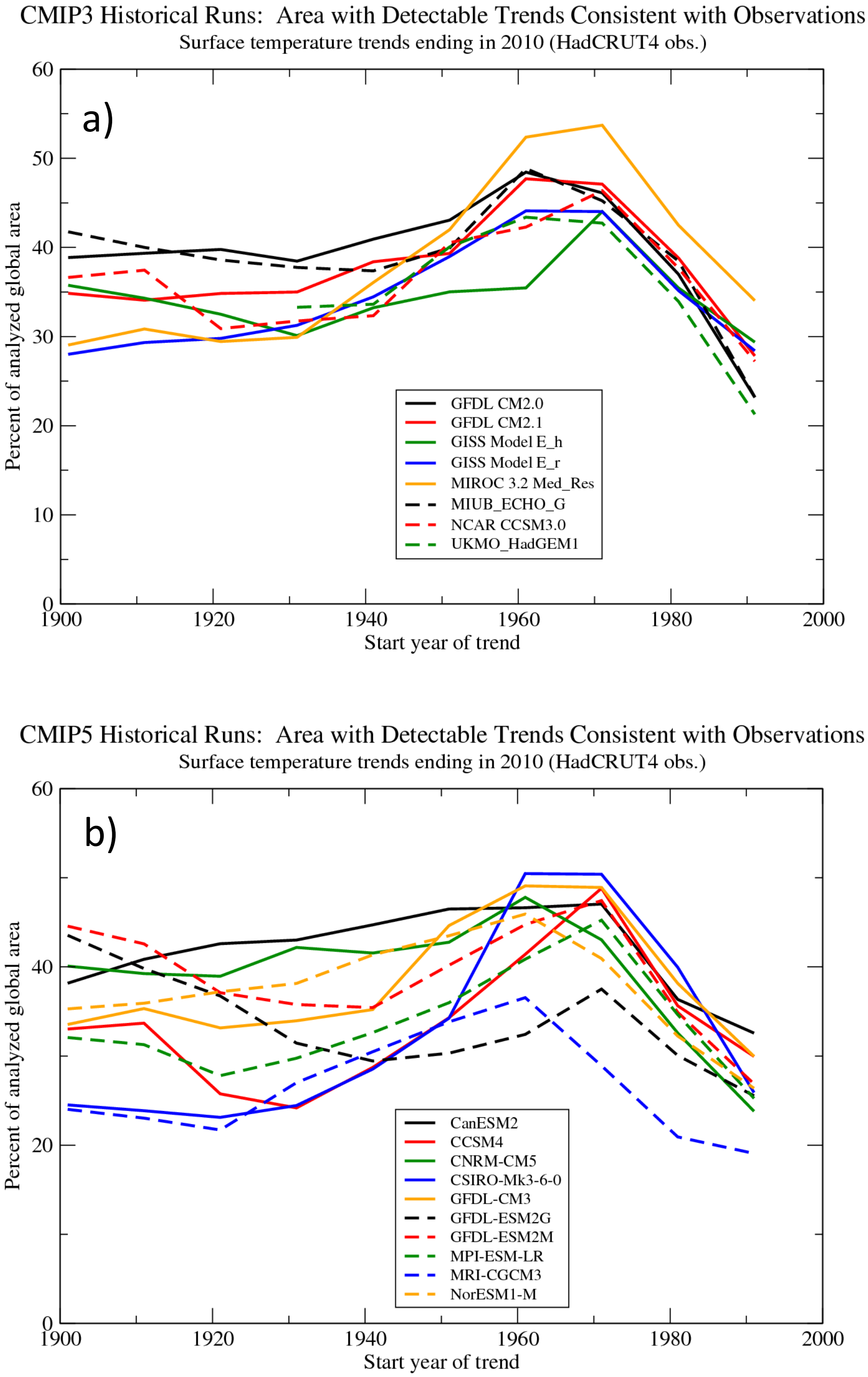

size=10, had4land25, temp_ANN
Obs trend (1981–2010), degC/100yr
Mod ens mn trend
Model ens. mean: 7 cat
a) Obs
b) Volc
c) Volc
-40 -4 -2 -1.5 -1 -0.5 -0.25 0 0.25 0.5 1 1.5 2 4 40
3 1 -1 -3
d) Volc Cat-3 0.01
e) Volc Cat-2 0.01
f) Volc Cat-1 0
g) Volc Cat0 0.42
Cooling - detec and more than simu
Cooling - detec and consistent
Cooling - detec but less than simu
No detectable change
h) Volc Cat+1 0.04
i) Volc Cat+2 0.45
j) Volc Cat+3 0.08
k) Volc Cat+1,+2,+3 0.54
Warming - detec but less than simu
Warming - detec and consistent
Warming - detec and more than simu
Warming or detectable
l) Volc Cat+2,+3 0.49
m) Volc Cat-2,0,+2 0.73
n) Volc Cat-3,-1,0,+1,+3 0.26
Detectable warming cons or gt simu
Obs and simu change are consis
Obs and simu change are incons
0 0.2 0.4 0.6 0.8 1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
80°N 40°N 0° 40°S 80°S
0° 100°E 180°W 80°W

Fig. 14



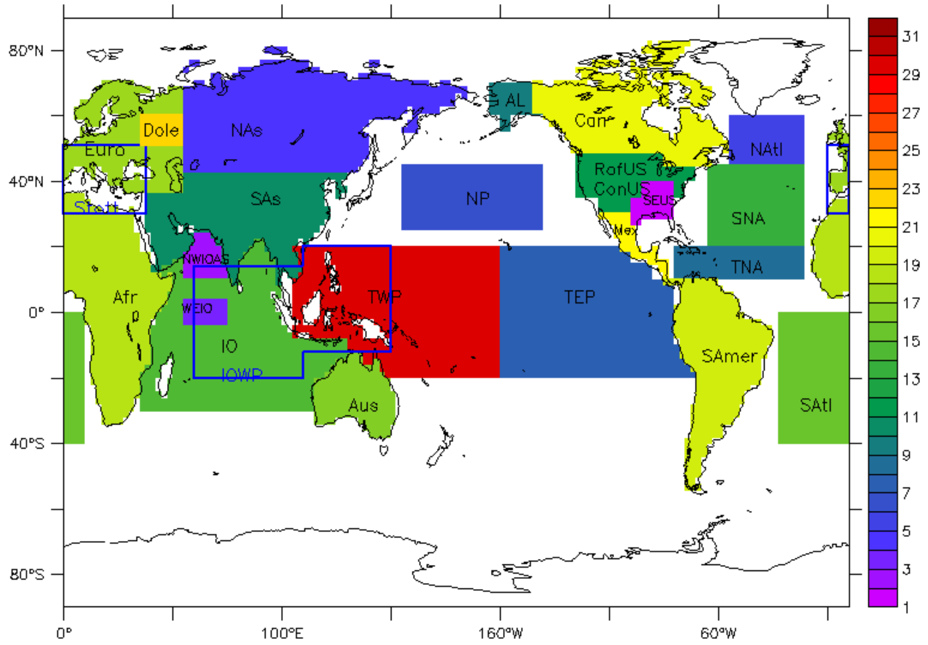
CMIP3 Historical Runs: Area with Detectable Trends Consistent with Observations

Surface temperature trends ending in 2010 (HadCRUT4 obs.)

CMIP5 Historical Runs: Area with Detectable Trends Consistent with Observations

Surface temperature trends ending in 2010 (HadCRUT4 obs.)

Fig. 15


80°N
40°N
0°
40°S
80°S
0°
100°E
160°W
60°W
Euro
Dole
NAs
Stott
SAs
AL
Can
NAtl
RofUS
ConUS
SEUS
Mex
NP
SNA
NWIOAS
TNA
Afr
WEIO
TWP
TEP
IO
IOWP
SAmer
Aus
SAtl
31
29
27
25
23
21
19
17
15
13
11
9
7
5
3
1